HOW 'GOOD' IS URBAN BUS ELECTRONIC FARE COLLECTION DATA?

HOFMANN, Markus; Dept. Of Informatics, Institute of Technology Blanchardstown, Dublin, Ireland; markus.hofmann@itb.ie

WILSON,Simon; Dept. Statistics, Trinity College Dublin, Ireland; swilson@tcd.ie

WHITE, Peter; Dept. of Transport Studies, School of Architecture and the Built Environment, University of Westminster, London, UK; whitep1@westminster.ac.uk

ABSTRACT

In recent years there is clear trend towards the utilisation of Electronic Fare Collection (EFC) data. The automatic storage of data supported the decision making process in many other industries since the early 90s. The analysis are often complex and exceed the creation of revenue related reports which was one of the original purposes of EFC systems. However, very little research has been carried out on the quality of such data especially in systems where human involvement is part of the recording process. Many European systems require the interaction of the bus driver to identify the bus stop when passengers board. Although more and more systems are nowadays equipped with Automatic Vehicle Location (AVL) systems the majority of operators still have to operate without such technological advances.

This paper analyses EFC data from a network where AVL data was not available. Three different aspects are investigated: Is the entered location of the bus stop correct? Is there a fatigue factor of bus drivers throughout the day?

Furthermore a Monte Carlo simulation demonstrated to what extent wrongly recorded data influences EFC analysis results.

The results of this paper are promising as the conclusion with regard to data quality is that although the recorded data is not complete it is accurate.

Keywords: Public transport, data quality, electronic fare collection data

INTRODUCTION

Results are only as good as the data. However, knowing the shortcomings and strength of any data set will contribute to the results of the analysis. EFC data are used more and more frequently in many academic and commercial public transport projects. Various data attributes are automatically recorded when passengers board a public transport vehicle. The

main focus of this paper is the quality of the data. This is a common concern among the research community, particularly as often the bus driver has a certain influence on the recorded data. In many transport networks it is the bus driver's responsibility to provide the system with information such as which bus stop the public transport vehicle is currently serving and therefore is also in charge of identifying the correct location of passenger boardings. This section shows three different analyses that provide results with regard to the reliability of the bus driver and his/her input. It further discusses the real impact on the measured data quality. The first approach analyses the frequency of recorded bus stops. The second approach analyses odd boarding distributions on the premise that more people board at busy bus stops. The third approach focuses on the arrival time of the vehicle at the bus stop on the premise that the arrival times on bus stops should not be the same for a number of bus stops. The last approach uses a Monte Carlo simulation that assesses how small errors in data would impact the result of a heuristic transfer journey classification algorithm.

BACKGROUND

Nowadays, many research and commercially orientated public transport projects use EFC data to generate various types of analyses (e.g. Barry et al. (2002), Zhao (2004), Wilson et al. (2005), Trepanier et al. (2007), Hofmann et al. (2009)). EFC data is automatically recorded when certain events occur. Such events could be a bus leaving the depot, a bus passes a bus stop or when passengers board the bus. However, unless buses work with an integrated APS it is the responsibility of the bus driver to record these events. This section analyses the reliability of bus drivers and the impact of this reliability on the overall quality of EFC data.

What Defines Data to be 'Good'

The main relevant dimensions for this analysis are completeness and accuracy. For example, when a study focuses on passenger boardings then all passenger boardings have to be represented by the data source (completeness) and all records have to reflect the correct data values (accuracy). The following analyses will demonstrate how the data were tested for completeness and accuracy. Perfect datasets are rare due to human or machine errors, however, knowing to what extend a dataset is incomplete still allows it to be used for analysis.

Data quality can be categorised into four sections (Strong *et al.*, 1997, Liebscher, 2005):

- Contextual Data Quality - relevancy, value added, timeliness, completeness and amount of data
- Intrinsic Data Quality - accuracy, objectivity, believability and reputation
- Accessibility of Data - accessibility and access security
- Representation of Data - interpretability, eases of understanding, concise representation and consistent representation.

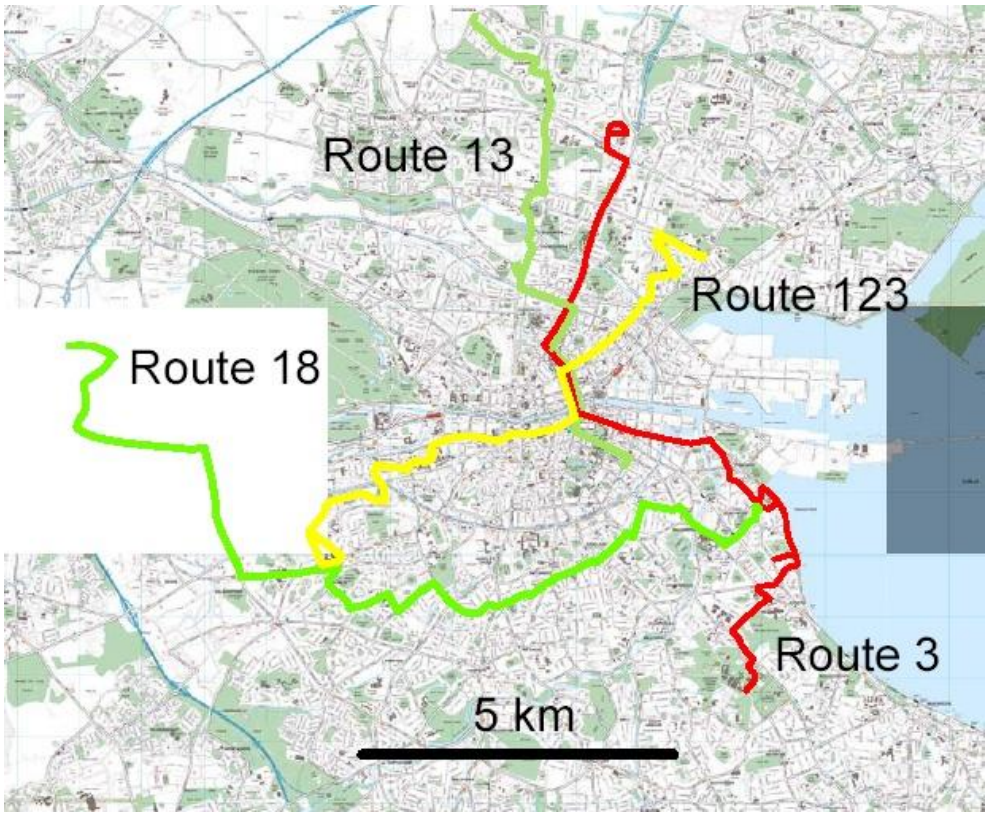
ANALYSIS METHODOLOGY

Four different approaches were used to test the integrity and quality of the data. The aim was to investigate the bus driver's consistency in recording the current location. The following three approaches were used:

- Frequency Tables
- Odd Boarding/Alighting
- Time Series
- Monte Carlo Simulation

Four different types of routes were subject to the analysis. The different characteristics of each route can be seen in FIGURE 1. It further shows the number of bus journeys carried out on each route for one particular day split into inbound and outbound journeys. The variability of changing numbers of bus journeys per day is very small and can therefore be disregarded.

The following sections will introduce the applied analysis methodologies.



Route Number	Number of Bus Journeys per Day		Type of Route
	Outbound	***Inbound***	
18	64	70	Orbital (West East)
13	59	51	North – City Centre
123	163	180	West – City Centre
3	95	87	Cross City Centre – North - South

FIGURE 1: Layout and Journeys of Analysed Routes

Frequency Tables

The first approach focuses on boarding records to analyse how diligently the bus driver carried out the task of recording the bus' location. This aspect of the analysis serves to analyse how exact this task has been carried out. The initial assumption of the approach was that each bus stop on a particular route should be recorded as often as any of the remaining bus stops. However, the location is only recorded in case the bus actually stops and passengers board. It is therefore expected that a certain pattern occurs which can be repeatedly produced analysing different bus routes with different route characteristics and different bus drivers. A smaller dataset only including data from a Wednesday in October 1999 was extracted for the analysis. A cross-tabulation was based on the dataset with the aim to obtain frequency numbers for each route's bus stops of that particular day. The data were then grouped into bus routes and direction of the journey. This format favoured the generation of bar charts which were then compared. A random selection of charts will be displayed and discussed.

Odd Boarding/Alighting

This approach focuses mainly on the boarding patterns of passengers. It is assumed that passengers are more likely to board at main bus stops such as shopping malls, main streets or multi modal transfer nodes. A randomly chosen set of bus stops are identified and explored with the main aim to find odd boarding patterns. Odd boarding patterns could be considered as patterns where passengers boarded unexpectedly which could lead to the assumption that the bus driver keyed in the wrong bus stop or forgot to indicate the correct bus stop. For example, if bus stop 5 on route X is the closest bus stop to the pedestrian zone then it is expected that more people board than on a bus stop outside the city centre (at least for an outbound journey). Main bus stops will be selected of which the researcher knows that a boarding pattern above the average is to be expected. This approach further focuses on bus stops where the route intersects with another mode of transport such as light rail or metro. These transfer nodes are expected to have a larger passenger boarding pattern than the previous or following bus stop. FIGURE 2 shows this scenario when the light rail intersects with the bus route at Stop 3 and Station 1. It is therefore expected to have a higher boarding pattern at bus stops 3 and 5 than at bus stops 4 and 6.

Time Series

The time series approach is the final test that is applied in assessing the quality of EFC data with regard to the influence of the bus drivers on data quality. The assumption is that the difference of bus stop arrival times of two bus stops cannot be less than a pre-defined period of time. This time frame depends on the network and also whether the bus stops were recorded in the city centre where stops are closer together than in suburban areas. Comparing the bus stop arrival times and calculating their differences can therefore provide an indicator of the data quality. A variability study strengthens this analysis.

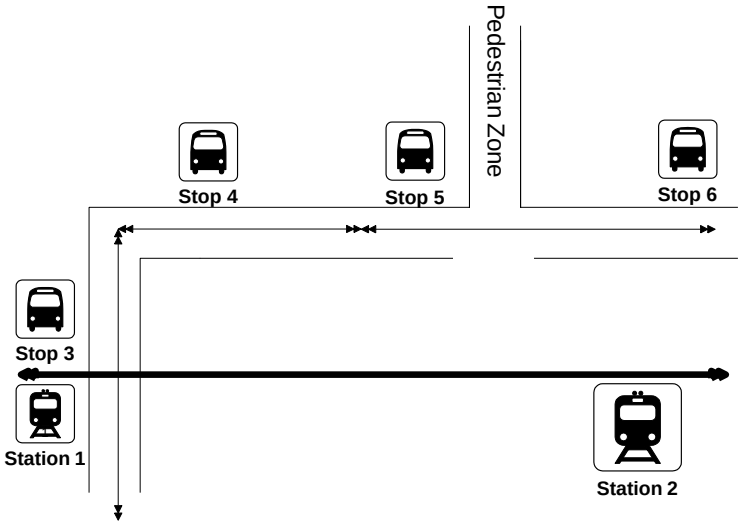


FIGURE 2: Sample Layout of a Transport Network Section

Monte Carlo Simulation

A Monte Carlo simulation was applied to a sample of the results to identify how a transfer journey identification algorithm (Hofmann, 2009) would perform if an error is randomly assigned to various attributes. Such an analysis aims to determine the robustness of the algorithm by testing its performance after randomly introducing a certain error percentage in the data files. This introduction of error could easily resemble the most common EFC errors. The simulation focused on the results of one day (Wednesday in April 1999). A total of 91,080 boarding records were stored for this particular day. We look at how the results of the transfer journey analysis change as we add error to the dataset.

The aim was to create datasets where 5%, 10%, 15%, 20%, and 25% of the value of one attribute were in error. The error was introduced by assigning a random valid value to that attribute. A C++ program randomly assigned the errors to the source file which was later used by the transfer journey identification algorithm. Three attributes that the transfer journey algorithm uses were looked at: Time of Boarding, Ticket ID and Ticket Type ID. The setup created 100 files for each attribute for each of the given error percentages. For each attribute and percentage error, 100 data sets were simulated and the transfer journey algorithm ran. The difference in the result with that of the original file is computed as percentage. This resulted in 500 files for each test. Although a larger number of iterations would strengthen this study due to the runtime of the transfer journey algorithm it was decided that 100 iterations is sufficient to analyse the robustness of the algorithm with regard to the most common data errors. The following paragraphs present the findings of the Monte Carlo simulation.

For each simulated file we measure the percentage of records that were classified differently (as transfer journey of non transfer journey) from the original file. We call this the classification error. This can also be expressed as

$$\epsilon = \frac{100}{n}\left(|TJ_{true} \cap (not\ TJ^*)| + |notTJ_{true} \cap TJ^*|\right) \quad (1)$$

where *n* is the number of records, *TJ* are identified transfer journeys in the original dataset and TJ^* are transfer journeys in the simulated file.
After the original data file was randomly changed and the transfer journey algorithm was executed it was necessary to analyse the results. The aim was to identify how many transfer journey pairs were matching when comparing the original result with the error induced result. Therefore false negative and false positive results in the Monte Carlo files were not included as correct transfer journey pairs. The false positive and false negative pairs could have occurred by randomly assigning a value to an attribute which would then lead by chance to a transfer journey pair.
Time Attribute: The Time attribute stored the time of boarding of the passenger. The value of this parameter was changed by randomly selecting 5%, 10%, 15%, 20%, and 25% of the records and then a random assignment of a value that was within the range of possible values. The range of possible values therefore reached from 00:00 to 23:59. Thus, time could only be randomly changed to another time but not to any random number. The purpose of changing the value of this variable was to simulate some of the most common EFC errors such as incorrect sign on, wrong time due to lack of maintenance and read-write errors.

RESULTS

Frequency Tables

The frequency tables were created for each bus route. The dataset consists of a grouped count of bus stops. The bus stops are in spatial sequential order for inbound and outbound journeys. As shown in FIGURE 3 the bus stop ID's (shown on the x-axis) are unique throughout each route hence the difference in bus stop numbers for the inbound and outbound journeys. For example bus stop ID '80' is the corresponding bus stop of ID '20' in the opposite direction.

The y-axis shows how often each particular bus stop was recorded throughout the day. The recording of the first bus stop is simultaneously the total number of journeys that were dispatched on the particular route and its direction (throughout one day). FIGURE 4 shows the distributions of recorded bus stops for the four routes used in this study. As expected each chart shows a reoccurring pattern for each route. The recording of the first bus stop is mandatory for the bus driver to start the recording of the journey. After recording the highest number of bus stops at the beginning of the journey the number of recorded bus stops generally drops slightly before levelling off for a few stops before consistently decreasing over the last few stops. The data show that the bus driver records less bus stops as the route approaches the final stops. This may reflect the fact that fewer passengers board during the last few stops as there is no need to board for only two or three bus stops. However, it was expected that more people exit the bus when approaching city centre which would imply that the bus had to stop more often than in suburban areas. This further implies that the bus driver only identifies the location of the bus when passengers are boarding. This may become a problem when calculating arrival times which will be further elaborated in Section 7.5 because of missing stage records towards the end of each journey as they are not as

frequent as they were at the beginning. It is also interesting that the bar charts consist of a certain amount of symmetry with regard to inbound and outbound journeys. This further assumes consistency of the bus driver's location recording pattern. Furthermore it seems that the number of outbound journeys marginally exceeds the number of inbound journeys. It is noteworthy that this pattern is not dependent on the time of the day. Various analyses have shown that peak time and off-peak time show the same characteristics with regard to the patterns shown in FIGURE 4. It could therefore be argued that the drop of recorded bus stops in the last section of the journey is not related to fatigue of the bus driver. Some graphs show a second peak during the journey recordings. These are popular bus stops or suburban areas where more passengers are expected to board. There are three ways to remedy that not all bus stops are recorded. The first method is to automate the recording using an AVL system that is integrated with the EFC system. The second option would be to make the bus drivers to press the button when passing the bus stop regardless whether passengers board or alight. The last option would require a statistical model that may be able to infer the arrival times of buses at bus stops based on historical data.

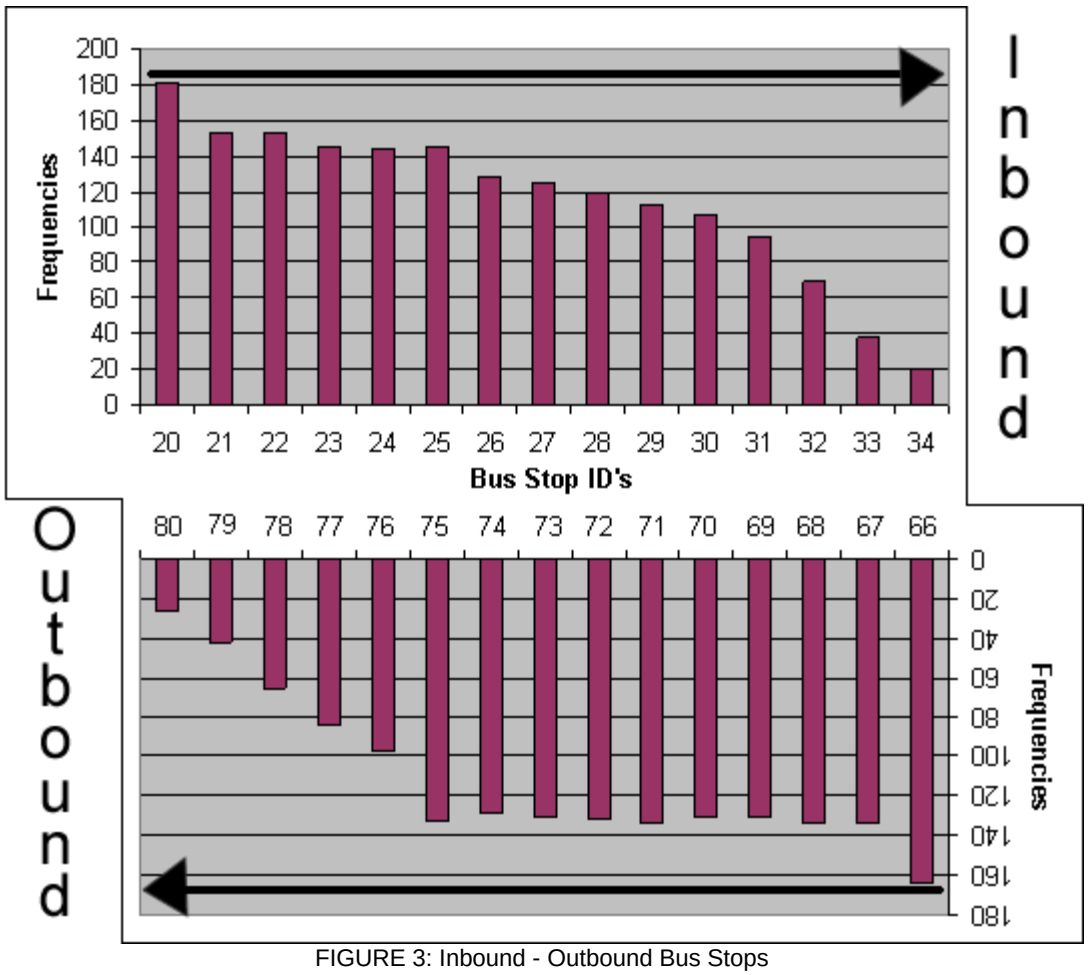
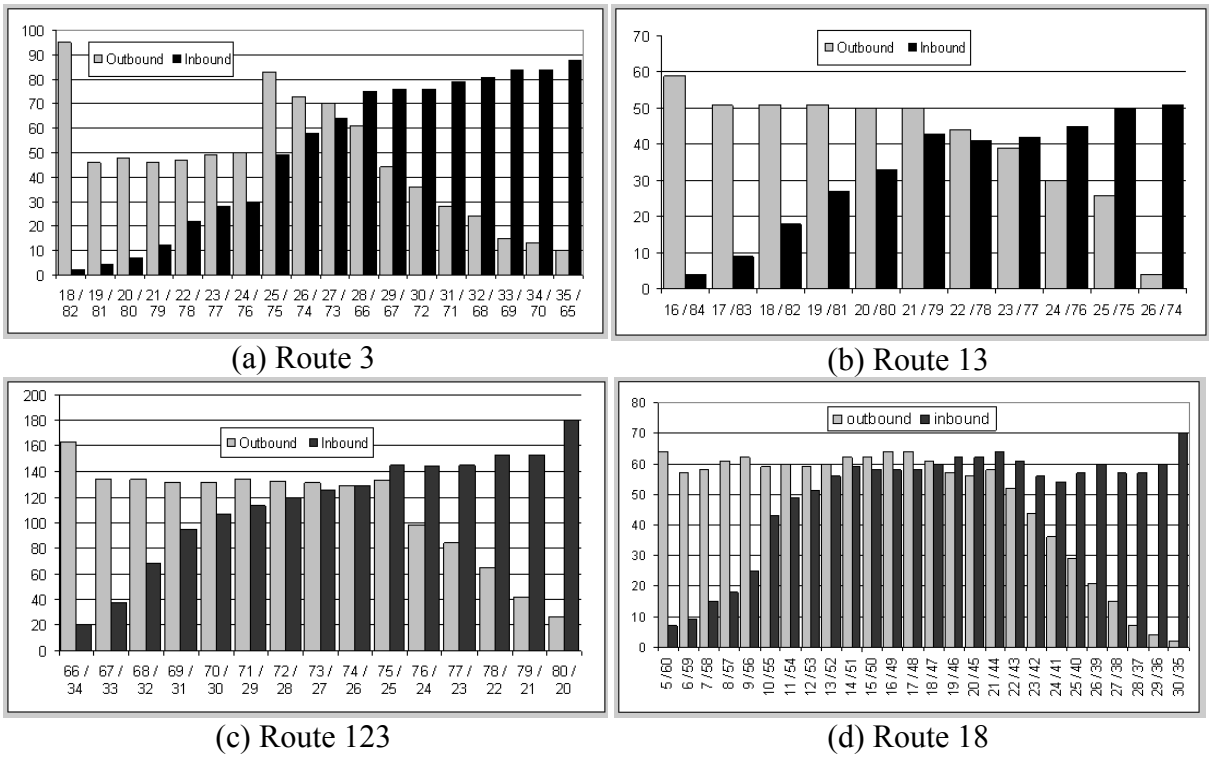


FIGURE 3: Inbound - Outbound Bus Stops



(a) Route 3 (b) Route 13

(c) Route 123 (d) Route 18

FIGURE 4: Frequency of Recorded Bus Stops

Odd Boarding

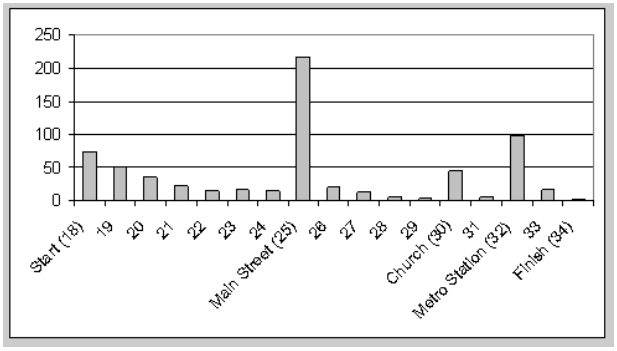
A frequency table was created showing the number of passengers that boarded on one particular day (Wednesday in October 1999). A series of bar charts then visualise the boarding patterns along a certain route. TABLE 2 shows the different routes which were subject to the analysis and their total recorded passengers throughout the day.

TABLE 2: Number of Passenger Boardings (Outbound/Inbound)

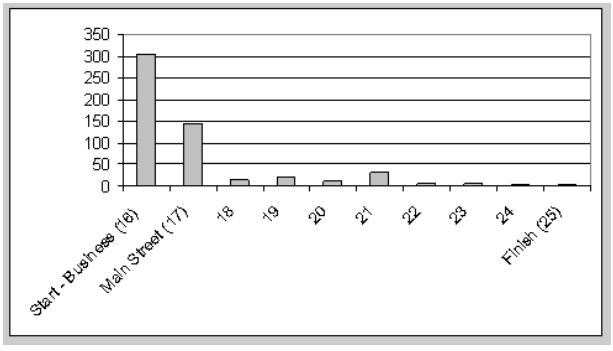
Route ID	Outbound	Inbound
3	649	583
13	594	542
123	939	1051
18	809	816

The table differentiates between boardings on outbound and inbound journeys. The variability between the day used in this study and other days is minimal and can therefore be disregarded. The approaching symmetry between total numbers of boardings of outbound and inbound journeys led to the decision that the direction of the journey with regard to passenger boarding numbers can also be disregarded. FIGURE 5 shows the results of the analysis. Each bar chart presents the boarding pattern of one route. The x-axis shows the identifier of the bus stop in sequential spatial order and also labels the stops where an increased boarding pattern occurred. The charts show an increase in passenger boarding at demographic landmarks such as shopping streets, churches, multi modal transfer nodes,

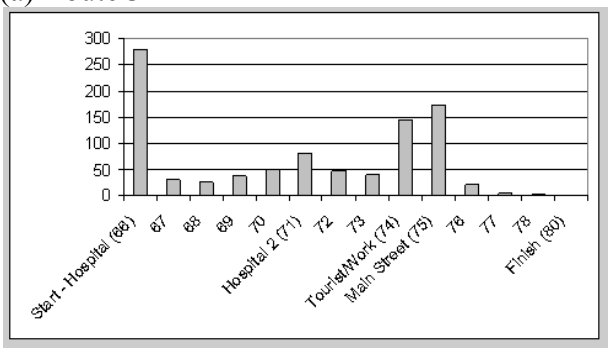
hospitals, suburban centres and schools. Considering the results it can be argued that the bus driver identified the correct bus stop when passengers boarded at these landmarks.



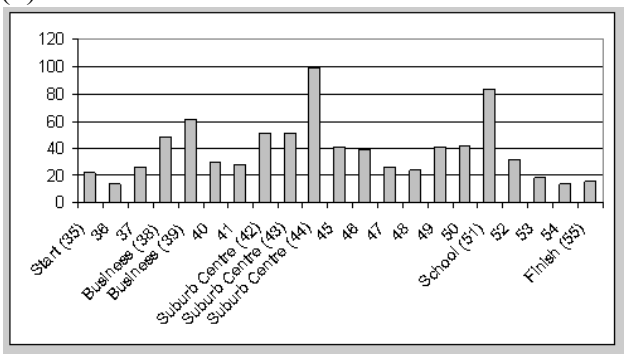
(a) Route 3



(b) Route 13



(c) Route 123



(d) Route 18

FIGURE 5: Frequency and Location of Recorded Bus Stops

Time Series

Unfortunately, the time recorded with each boarding only includes hours and minutes in a 24 hour format but does not contain seconds. This prohibits an exact analysis because in theory the bus could serve two bus stops within the same time stamp (time recorded to nearest minute), particularly in the city centre where the distances between bus stops are short or along dedicated bus corridors where the speed of the bus would be increased. It was decided to show a graphical representation of arrival times of several buses at each bus stop. Different peak/off peak times and directions (inbound/outbound) were chosen. FIGURE 6 show the different scenarios. The horizontal axis shows the sequential bus stop IDs in a non-spatial manner. The vertical axis shows the time of arrival. Each line within the line charts indicates a bus that serves a number of bus stops. Horizontal bars connecting two nodes indicate that the time recorded for the two bus stops was identical.

FIGURE 6(a) shows the arrival times of several buses serving the Route 123 in outbound direction throughout the morning peak period. The arrival times of buses at the various bus stops are different most of the time. However, it was noticed that the time recorded towards the end of the route often remains the same which would indicate that the bus driver did not identify the correct bus stop. It could be argued that only passengers with a magnetic strip card boarded that close to the final stop and the identification of the correct bus stop was therefore not necessary. FIGURE 6(b) shows the arrival times of several buses serving

Route 13 in outbound direction throughout the afternoon off peak period. The same explanation as given above applies to this graph. Apart from a few times throughout the route the recorded time difference over the period of the service was different, indicating that the bus driver keyed in the correct bus stops. This graph also shows that there is no problem with regard to bunching. The frequency of the service is regular and the headway is spread evenly. FIGURE 6(c) shows the arrival times of several buses serving Route 18 in inbound direction throughout the evening peak period. Again, only a few bus stops were recorded with the identical time to the previous bus stop. Although this is not related to data quality it has to be pointed out that the graph favours the presentation of bunching and headway. FIGURE 6(d) shows the arrival times of several buses serving the Route 3 in inbound direction throughout the morning peak period. For some reason some of the bus services did not collect data for the last three to five bus stops. Although this pattern was noticed to a certain extent throughout the other routes it was not as drastic as identified with route 3. There is no explanation for this pattern. It could have been that some of the routes were scheduled not to provide a service after bus stop '75' (O'Connell Street – Main City Centre Street). The following points could be identified when analysing the sequential time differences of arrival times:

- The value '0' (less than 1 minute between bus stops) appeared 38% (Route 3), 22% (Route 13), 27% (Route 18) and 29% (Route 123). This indicates that there is considerable variability between the different routes;
- The Median lies at 2.00 minutes;
- There is a trend that inbound journeys have more '0' values than outbound journeys.

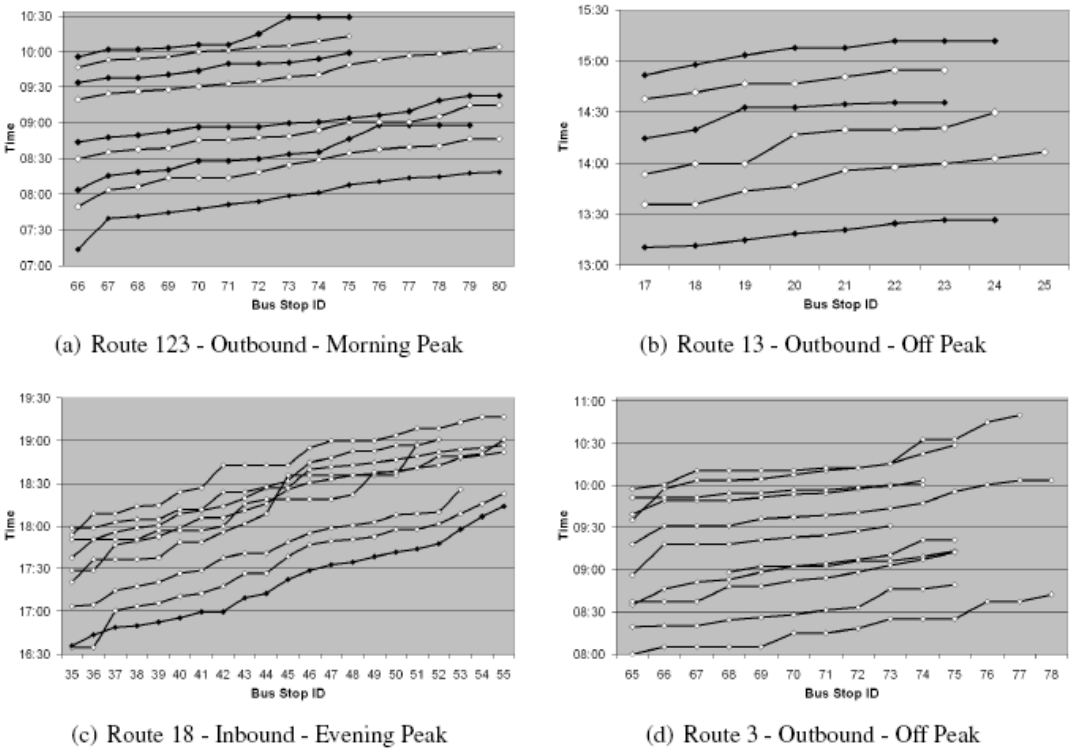


FIGURE 6: Arrival Time at Bus Stop

Validation of the Results Using a Monte Carlo Simulation

Error! Reference source not found. shows the descriptive statistics of the Monte Carlo simulation for the various induced error percentages of the Time attribute. The error in identifying transfer journeys runs almost linear in relation to the randomly introduced error (see **Error! Reference source not found.**7(a)).

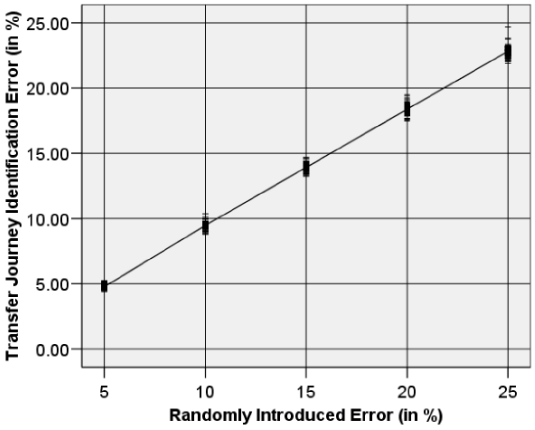
TABLE 3: Summary of Classification Error (as %) using Monte Carlo Simulation

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Time Attribute Simulation Study							
5% Error	100	.86	4.40	5.26	4.78	.1798	.0323
10% Error	100	1.57	8.80	10.37	9.48	.2651	.0703
15% Error	100	1.44	13.25	14.69	13.93	.3089	.0954
20% Error	100	2.00	17.49	19.49	18.38	.3968	.1575
25% Error	100	2.78	21.91	24.69	22.82	.4014	.1611
Time Attribute (60 Minute Range) Simulation Study							
5% Error	100	.28	.02	.30	.16	.0463	.0021
10% Error	100	.25	.20	.45	.33	.0573	.0033
15% Error	100	.33	.34	.67	.51	.0674	.0045
20% Error	100	.32	.55	.87	.70	.0684	.0047
25% Error	100	.39	.66	1.05	.90	.0657	.0043
Time Attribute (15 Minute Range) Simulation Study							
5% Error	100	.25	.00	.25	.07	.0554	.0031
10% Error	100	.43	.00	.43	.13	.0950	.0090
15% Error	100	.45	.00	.45	.19	.1087	.0118
20% Error	100	.59	.00	.59	.30	.1118	.0125
25% Error	100	.60	.12	.72	.40	.1299	.0169
Ticket ID Attribute Simulation Study							
5% Error	100	.95	7.68	8.63	8.20	.179	.0321
10% Error	100	1.08	15.70	16.78	16.22	.227	.0516
15% Error	100	1.46	23.46	24.92	24.11	.272	.0741
20% Error	100	2.12	30.81	32.93	31.82	.338	.1143
25% Error	100	1.53	38.93	40.46	39.54	.261	.0682
Ticket Type Attribute Simulation Study							
5% Error	100	.79	9.04	9.82	9.40	.161	.0261
10% Error	100	1.12	17.93	19.05	18.56	.211	.0445
15% Error	100	1.68	26.75	28.43	27.52	.266	.0705
20% Error	100	2.38	35.57	37.95	36.28	.322	.1037
25% Error	100	2.82	44.61	47.43	45.02	.316	.1001

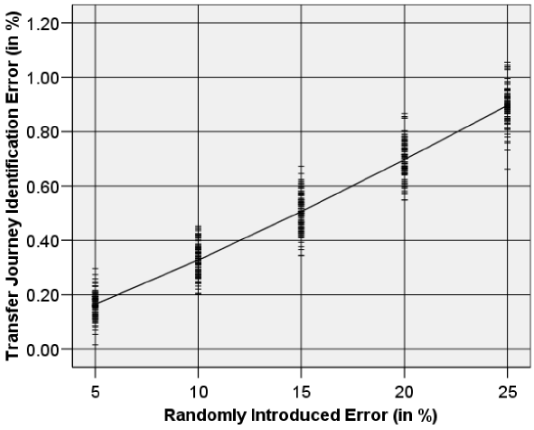
This analysis lead to a more specific Monte Carlo simulation in the sense that the range of possible values was decreased. Two more simulations were defined: the first simulation included data files which time values were randomly changed to values in a range of -15 to 15 minutes; the second simulation used a range of ±60 minutes.

Error! Reference source not found. and FIGURE 7(b) show the results of the ±60 minute simulation. This Monte Carlo simulation shows that the algorithm performs very well when

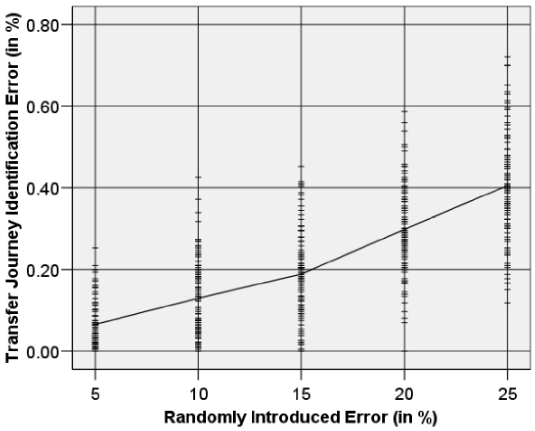
small errors with regard to the time attribute are introduced. Even a 25% error rate only results in an average 0.9% error of the transfer classification algorithm.



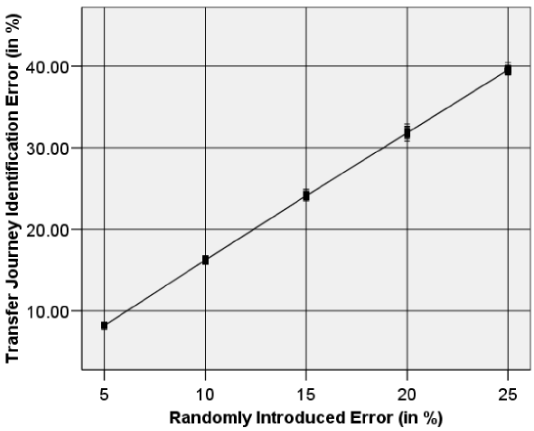
(a) Time Attribute - Full Range



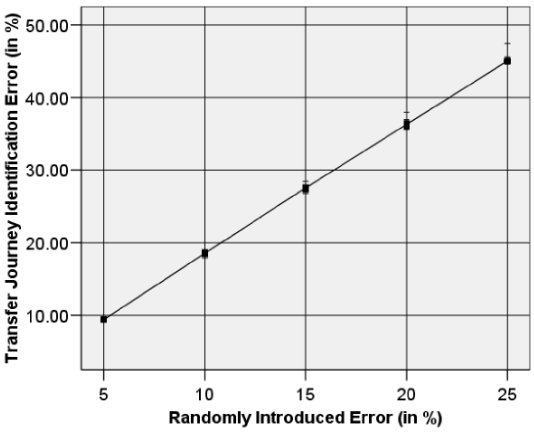
(b) Time Attribute - 60 Minute Range



(c) Time Attribute - 15 Minute Range



(d) Ticket ID Attribute



(e) Ticket Type Attribute

FIGURE 7: Monte Carlo Simulation Results

Error! Reference source not found. and FIGURE 7(c) show the results of the Monte Carlo simulation of the time attribute with introduced errors of ±15 minutes. The average caused transfer identification error of the algorithm is at 25% error only 0.4%. It can therefore be stated that small EFC errors with regard to the time attribute can be neglected.

In summary, we believe that small errors for the time attribute have no significant impact on the results produced by the transfer journey algorithm. This is mainly due to the selection of m = 90 minutes. False negatives and false positives can occur when the error changes the time in such a manner that the time difference of the two boardings is around 90 minutes. The Monte Carlo simulation clearly showed evidence that supports this statement.

Ticket ID: This attribute stored the Ticket ID which is in combination with the ticket type a unique number which can be used to identify individual passengers over a certain period of time. The value of this parameter was changed by randomly selecting 5%, 10%, 15%, 20%, and 25% of the records and then a random assignment of a value that was within the range of possible values.

Error! Reference source not found. shows the descriptive statistics of the Monte Carlo simulation for the various induced error percentages of the Ticket ID attribute. The graphical representation of this simulation can be seen in FIGURE 7(d). Randomly introduced errors cause a high transfer journey identification error. This had to be expected as changing Ticket ID assigns the boarding record to a 'new' passenger.

Ticket Type ID: This attribute stored the Ticket Type ID which is in combination with the Ticket ID a unique number which can be used to identify individual passengers over a certain period of time. The value of this parameter was changed by randomly selecting 5%, 10%, 15%, 20%, and 25% of the records and then a random assignment of a value that was within the range of possible values.

Error! Reference source not found. shows the descriptive statistics of the Monte Carlo simulation for the various induced error percentages of the Ticket Type ID attribute. The graphical representation of this simulation can be seen in FIGURE 7(e). Randomly introduced errors cause a high transfer journey identification error. This had to be expected as changing Ticket Type ID assigns the boarding record to a 'new' passenger that may or may not already exist.

The various Monte Carlo simulations showed that recorded error in the attributes Ticket ID, Ticket Type ID and Time result in high transfer journey identification errors. However, smaller changes in time result in error rates that can be almost neglected. Even when 25% of all time values were changed by either up to 15 or 60 minutes the maximum transfer journey identification error was only 1.05%. The algorithm therefore performs very well with regard to small errors of the time attribute which is also one of the most likely incorrect attributes as system time is not set centrally. Errors in Ticket Type ID or Ticket ID on the other hand result in a larger error because the boarding record is attributed to a different passenger when either of these values change.

DISCUSSION & CONCLUSION

All four approaches presented in this paper show that the bus driver interaction has a consistent pattern with regard to recording the bus location and thus the boarding time. There are error margins which have to be considered when interpreting the results obtained

from EFC data analysis. The four routes that served as subjects of this analysis represent average radial and orbital routes.

In summary, the interaction of the bus driver as location indicator has certainly an impact on the data quality. However, the main recordings are correct and it is debateable whether the error margin would bias the results considerably. Even if an actual in-vehicle time of 35 minutes is inferred as 32 or 38 minutes it is still better than having no identification of such a performance measure at all. Taking the data quality measurements introduced by Strong et al. (1997) this study can conclude that the accuracy of the data is, within a margin of error, acceptable. After having carried out the analysis it was concluded that the data quality of the recorded records is representative. However, it was further found that the records that were not recorded (e.g.in case no one boarded) but could have been recorded have a great impact when focusing on stage level analyses. Therefore the data quality measurement 'completeness' introduced by Strong et al. (1997) is not entirely fulfilled. This especially applies to the bus stop records as demonstrated above. In theory such missing records could be inferred. For example, missing stage records could be inferred using a Geographic Information System (GIS) and the data of the recorded stages in order to estimate the arrival time of the missing stages by considering time of previous and subsequent record, average speed and distance travelled. However, this was not further pursued as no GIS of the bus network was available

The Monte Carlo simulation showed that even when a certain degree of error exists and therefore reduces the quality of the data it can still be used for analyses. The importance lies in knowing the degree of data quality prior to its usage.

REFERENCES

Barry J.J., R. Newhouser, A. Rahbee, and S. Sayeda. Origin and Destination Estimation in New York City with Automated Fare System Data. Transportation Planning and Analysis, Transportation Research Record, 1817:183 – 187, 2002.

Hofmann M., S. Wilson and P. White. Automated Identification of Linked Trips at Trip Level Using Electronic Fare Collection Data. Washington , D.C. , USA , 2009. Transportation Research Board (CD-ROM).

Hofmann M., S. Wilson and P. White. Automated Identification of Linked Trips at Trip Level Using Electronic Fare Collection Data. Washington, D.C. , USA , 2009. Transportation Research Board (CD-ROM).

Liebscher, R. Informational support of public transport operations in exceptional situations. European Transport Conference, Strasbourg, France, 2005.

Plotnikov, V. An Analysis of Fare Collection Costs on Heavy Rail and Bus Systems in the U.S. PhD Thesis, Virginia Polytechnic Institute and State University, USA, 2001.

Strong D.M., Y.W. Lee, and R.Y. Wang. Data Quality in Context. Commun. ACM, 40(5): 103–110, 1997. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/253769.253804.

Trepanier, M, Tranchant, N., and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems, 11(1):1–14, 2007a.

Wilson N.H.M., J. Zhao, and A. Rahbee. Extending the Value of Automatic Data Collection Systems. Invited Presentation at the Annual Transportation Research Board Meeting, 2005.

Zhao, J. The Planning and Analysis Implications of Automated Data Collection Systems: Rail TransitOD Matrix Inference and Path Choice Modeling Examples. MSc Thesis, MIT – Massachusetts Institute of Technology, September 2004.